

Does An Integrated Medical Workstation Really Help Clinicians? - A Formal User Evaluation

E.M. van Mulligen^{1,2}, T. Timmers¹, and J.H. van Bommel¹

¹Dept of Medical Informatics, Erasmus University,

²University Hospital Dijkzigt,
Rotterdam, The Netherlands

A formal user evaluation of an integrated medical workstation for support of clinical data analysis is described. Twenty-six users participated in an assessment study that consisted of a self-instruction course, followed by an experiment in which six clinical data analysis problems had to be solved, and the assessment was concluded with an evaluation form. To facilitate an objective and quantitative assessment, the time spent to the course and to the problems, as well as completeness, correctness, number of solving attempts and persistence were measured. Not all problems were solved by all participants. From an analysis of the measurements, the following conclusions were drawn. (1) Clinicians correctly solved about 69% of all problems tackled. However, they skipped 33% of the problems. (2) Non-clinicians correctly solved 87% of the tackled problems and skipped only 10%. (3) The performance of users not experienced in clinical data analysis was raised to the level of those with clinical data analysis experience. (4) An inventory which can be used to direct future developments was made of the errors of interaction. This assessment study has contributed to gaining insight into the conceptual problems and practical bottlenecks clinicians have with clinical data analysis.

INTRODUCTION

During the last five years, much research was done in the area of developing integrated medical workstations. Many researches investigated the possibilities of new technologies as a basis for integrated workstations [1-3]. New paradigms have been presented for the physician's desktop that address the issues of task-orientation, modularity, connectivity [4] and communication abstraction [5,6]. Earlier, we presented our own work on a prototype medical workstation (MW2000) for support of clinical data analysis [7], and on a refinement and generalization of the prototype, resulting in the HERMES workstation [8,9].

Our aim is the development of an architecture that can be used for integrating existing applications

without modifications. This architecture uses a client-server model to connect applications through the network and provides a uniform user-interface layer around these applications. This interface also supports automatic translation of data between the different formats, and generation of commands or keystrokes for the applications. In addition, there is a dual data model that on the one hand offers users a medical view of all data available in the network, and on the other hand a data dictionary that corresponds with the database schemas of the different integrated databases [10,11].

User evaluation

The prototype MW2000 has been concluded with a user evaluation. The objective of this evaluation was twofold. Firstly, to test whether clinicians do benefit from an integrated workstation for their clinical data analysis and consequently, if it is worthwhile to continue in this direction. Secondly, to get a better insight into the problems users had with the prototype, in order to use this both to guide new developments and as a reference for future assessments.

Although evaluation studies have been conducted in the field of clinical decision support systems [12,13] and computer support, we do not know of any publication of a formal evaluation of integration technology. In general, evaluation studies typically answer questions such as: Is the system of good quality? Does it reason/behave appropriately? Most of those studies address the effects on patient care. Consequently, the type of measurements described in these studies is not directly useful for our evaluation. The evaluation described here does not focus on the use of the applications themselves, but rather on the quantification of the added value of integration itself.

Most physicians have their clinical data analyses done by others. Therefore, using the workstation cannot be compared with a current situation for physicians. We included biomedical researchers in the study both as a reference group for the physicians and to assess their performance with the system. Moreover, we divided physicians and biomedical

researchers into two groups according to experience with clinical data analysis.

In the following, the assessment study is described together with the results, and a discussion of the results. More details are presented in [14].

MATERIALS AND METHODS

Setup

As explained, the workstation provides such a big qualitative step forward regarding support to clinical data analysis, that no realistic comparison can be made with the conventional kind of support. We decided to assess the workstation along three main lines:

1. can clinicians learn to use it?
2. can they use it to solve typical clinical data analysis problems?
3. is their performance comparable to that of statisticians?

The participants were introduced to the assessment study through a written, self-instructing tutorial that explained the principles of the workstation and how to use the window-based user-interface. The tutorial also explained the different steps of data analysis in the workstation: formulation of a query to a database in the network, saving data locally as a data-set, applying a statistical test, and graphical presentation of the results. Two examples demonstrated these steps explicitly. The tutorial ended with two data analysis problems; the user was requested to solve these problems. If a user was not able to answer these problems, he/she was not invited to participate in the evaluation experiment. All users succeeded in solving these two problems.

Users and groups

Twenty-seven users (13 biomedical researchers and 14 clinicians) entered the study. All clinicians were recruited from the University Hospital Dijkzigt. The biomedical researchers were recruited from the Faculty of Medicine of the Erasmus University Rotterdam, the University Hospital Dijkzigt and from medical institutions outside the Erasmus University. One user followed the tutorial but was not able to participate in the study.

The remaining 26 participants were divided into two groups in two different and independent ways. Firstly, according to their involvement in patientcare: 8 clinicians (CI) and 18 non-clinicians (NCI). Secondly, for each individual user we examined his/her publications in the period of 1987

to the first half of 1992 as indexed in MedLine. The 13 users who published during that period as a primary author or who had at least one paper each year as co-author were considered as experienced in clinical data analysis (E). The thirteen remaining users were assigned to the non-experienced group (NE). Comparisons within each pair of the above four groups were done according to the measurements described below.

Test Problems

For the assessment, we used a database with data of the follow-up of colon cancer, collected in the University Hospital Leiden by Bloem and described in [15]. This database includes demographic variables and variables that specify the type and location of the tumor, the therapy and the time of survival and the Dukes stage as an indicator for the severity of the tumor.

Six clinical data analysis problems (I-VI) were defined: three problems (I-III) were presented together with an ordered list of all essential steps (stepwise problems) and three (IV-VI) without (open problems). Each stepwise problem had a similar open counterpart. The problems in each set were gradually becoming more complex: the last problem dealt with a statistical procedure not described in the introductory tutorial. Three problems (II,IV,V) consisted each of an A and B part, requiring identical analysis for each different subset of the data (e.g., A = males, B = females).

Measurements

The following measurements were obtained: time necessary for the tutorial, time necessary for solving the problems, correctness, completeness, and number of solving attempts. During the study, the workstation logged the user's actions and the responses of the workstation. These logging data were analyzed for completeness and correctness as compared with the obvious minimal, correct path to solving the problems.

Scores

From these measurements, the following scores were defined per user:

- S_1 : percentage of correctly completed problems
- S_2 : percentage of completed problems
- S_3 : percentage of answered problems
- S_4 : percentage of unanswered problems
- S_{1a} : percentage of correctly completed stepwise problems
- S_{1b} : percentage of correctly completed open problems
- C : fraction of correctly completed problems from all answered problems

Apparently, S_1 is a subset of S_2 which is a subset of S_3 . $S_3 + S_4 = 100\%$. S_1 is the percentage of the total number of problems; C is the percentage of the problems that were answered (S_3).

RESULTS

In Table 1, the mean measurements are presented for the different groups. Column T_c shows the average time per group spent on the introductory course. Although the workstation was new to all users and none of the users had previous experience with the type of user interface of the workstation, the course was completed in about two hours by almost all users. Column T_e gives the average time necessary to solve one problem. The average durations vary from 10 minutes to 13 minutes for solving one problem. This includes reading the problem, selecting the data, analyzing them, and printing graphical presentation(s).

The clinicians (CI) did have a lower score for both completeness and correctness than the non-clinicians (NCI). These differences are larger for the open questions ($S_{1\alpha}$) than for the stepwise questions ($S_{1\alpha}$).

P -values were computed to test for statistically significant differences between the groups. No significant differences were found for the measured durations for any of the groups. For clinicians, the S_1 ($p=0.04$), S_2 ($p=0.01$), S_3 ($p=0.01$), $S_{1\alpha}$ ($p=0.01$) and S_4 ($p=0.04$) are significantly different. Besides these, no other significant differences were found.

Clinicians (CI) have a higher score S_4 . This indicates that this group of users more often skips a problem than the non-medical users. The lower number of problems answered also contributes to a lower correctness (S_1) and lower completeness (S_2) score.

Table 1. Mean measurements per group of users

	#	T_c	T_e	S_1	S_2	S_3	$S_{1\alpha}$	$S_{1\beta}$	S_4	C	P_1	P_n
CI	8	136	13.0	.54	.64	.67	.69	.43	.33	.69	5.6	21.1
NCI	18	114	10.5	.81	.88	.90	.82	.81	.10	.87	1.8	9.1
E	13	125	9.9	.74	.82	.83	.82	.68	.17	.83	5.9	25.0
NE	13	116	12.6	.72	.79	.82	.73	.71	.18	.80	0	0.46

CI/NCI	=	involved in patient care/not involved
E/NE	=	research experience derived from MedLine/no experience
#	=	number of users
T_c	=	average time of course (min)
T_e	=	average time per question answered (min)
S_1	=	percentage of correct and complete questions
S_2	=	percentage of complete questions
S_3	=	percentage of answered questions
$S_{1\alpha}$	=	percentage of correct and complete questions in I-III
$S_{1\beta}$	=	percentage of correct and complete questions in IV-VI
S_4	=	percentage of unanswered questions
C	=	percentage of correctly answered questions of all questions with an answer (S_1 / S_3)
P_1	=	average number of publications MedLine 1987-1992, if first author,
P_n	=	average number of publications MedLine 1987-1992, if co-author.

Clinicians (CI) had more publications as a primary author indexed in MedLine in the period of 1987 until 1992 than non-clinicians (NCI). Six clinicians (of 8) are assigned to the group of experienced authors (E). In Table 2, the intersections of the clinicians (CI) and the experienced authors (E) are given. Note that non-clinicians with experience contribute substantially to the high score of the E group. From Table 2 it is also clear that the large

S_4 score for clinicians is mainly due to the clinicians with no experience.

The lower C score of clinicians is mainly due to a larger number of errors. In Table 3 these errors are shown for clinicians and non-clinicians. Most problems were due to wrong data selection and forgetting to save data in a data-set. Other problems

are due to lack of knowledge about properties and parameters of statistical tests.

Table 2. Measurements for intersections of both groups

Group	#	S ₁	S ₂	S ₃	S ₄	S _{1α}	S _{1β}	C
CI+E	6	.59	.70	.72	.28	.71	.50	.69
NCI+E	7	.87	.92	.92	.08	.93	.83	.94
CI+NE	2	.39	.44	.50	.50	.63	.20	.68
NCI+NE	11	.78	.85	.88	.12	.75	.80	.83

See for abbreviations Table 1.

Table 3. Number (% per user) of user errors during the assessment study

Type of Error	CI		NCI		Total
Error in data selection	40	(2.6)	24	(0.7)	64
Wrong variables for cross-table	13	(0.9)	17	(0.5)	30
More than 2 outcomes for chi-square test	24	(1.6)	18	(0.5)	42
No saving of data	45	(2.9)	71	(2.1)	116
No table for computing p-values	8	(0.5)	7	(0.2)	15
Variables of cross-table interchanged	7	(0.5)	4	(0.1)	11
No variables in cross-table selected	22	(1.4)	26	(0.8)	48
Wrong parameters specified for survival table	33	(2.1)	25	(0.7)	58
Total	192	(12.5)	192	(5.6)	384

CI = clinicians, involved in patient care
NCI = not involved in patient care

DISCUSSION

From the assessment study, it appears that the evaluation of a system such as the prototype workstation with its high userinteraction is very complex. We also realized that references in the literature on quantitative rather than qualitative assessments are very sparse if not totally absent. Yet, we were convinced, and still are, that progress

in the further development of these and similar systems can only be made when the evaluation of such systems is done in a quantitative and objective as possible manner.

From the time necessary to understand the operation of the workstation (± 2 hours), and the short average time per question (10-13 min), we conclude that the workstation offers clinicians an environment through which they can master the network and the interaction with different applications on different hosts.

Clinicians and non-clinicians are statistically significantly different for S1 (correctness), S2 (completeness), S3 (problems solved), and S4 (problems not answered). This was primarily caused by the fact that clinicians skipped 33% of the problems, whereas non-clinicians only skipped 10%. This large total of unanswered problems is mainly induced by clinicians without clinical data analysis experience.

For the correctly solved fraction of all problems tackled no statistically significant difference existed ($p=0.06$) between clinicians and non-clinicians. In contrast to the open problems, no differences were found for the stepwise problems between clinicians and non-clinicians. From the assessment it seems that clinicians are less inclined to repeat a test. Once they have seen how it can be solved in principle, the clinicians can equally well solve problems with the workstation as non-clinicians.

From the absence of statistically significant differences between users who are experienced and non-experienced in clinical data analysis, it appears that the workstation supports non-experienced users to an extent that they are capable of solving clinical data analysis problems as well as experienced users.

The evaluation study also brought insight into unexpected flaws in the design of the user interface (Table 3). The errors as registered in the logging file, show that the selection of data caused many problems. A second type of error that often occurred is the saving of data-sets locally. This additional step was included in the prototype to prevent unintentional storing of large data-sets locally. Clearly, clinicians and non-clinicians did not recognize this explicit step, probably because all other steps were performed automatically and without interference of the user. Consequently, it is felt that the support in the prototype workstation was not yet sufficient for medical users. In the new HERMES environment [8], this has all been improved.

Acknowledgements

We thank Prof.dr. A. Zwaveling of the University Hospital Leiden for permission to use the colon cancer database, and dr. Th. Stijnen of the Center of Epidemiology and Biostatistics of the Faculty of Medicine of the Erasmus University Rotterdam for advice on the analysis of the assessment data.

Reference

- [1] Degoulet P., Coignard F.C.J., Laurent M.C., Lucas L., Ben Said M., Meinzer H.P., Engelmann U., Springub A., Baud R., Scherrer J-R. The HELIOS European project on software engineering. In: Timmers T., Blum B.I., eds. *Proceedings of the IMIA Working Conference on Software Engineering in Medical Informatics*. Amsterdam, October 1990. North-Holland, Amsterdam. 1991:125-37.
- [2] Tang P.C., Annevelink J, et al. Physicians' workstations: integrated management for clinicians. In: Clayton P.D., eds. *Proceedings of the 15th Symposium on Computer Applications in Medical Care*. Washington DC. New McGraw-Hill, New York. 1991:596-73.
- [3] Johnson S.B., Clayton P.D. Fink D., Sengupta S., Shea S., Bourne P., Sideli R.V., Aguirre A., Cimino J.J., Hripcsak G., McCormack M., Hill C. Achievements in phase III of an integrated academic information management system. In: Lun K.C., Degoulet P., Piemme T.E., Rienhoff O., eds. *Proceedings of the Seventh World Congress on Medical Informatics*, Geneva, North-Holland, Amsterdam. 1992:117-23.
- [4] Greenes R.A. "Locators", "Constructors" and "Trackers": meta-level tools for supporting health-care professional information needs in a distributed computing milieu. In: Lun K.C., Degoulet P., Piemme T.E., Rienhoff O., eds. *Proceedings of the Seventh World Congress on Medical Informatics*, Geneva. North-Holland, Amsterdam, 1992:2-7.
- [5] Barsalou T., Wiederhold G. Knowledge-directed mediation between application objects and base data. *Proceedings of the Working Conference on Data and Knowledge Base Integration*. October 1989.
- [6] Barsalou T. View objects for relational databases. PhD thesis, Medical Information Sciences Program, Stanford University, 1990.
- [7] Van Mulligen E.M., Timmers T., Van den Heuvel F. A framework for uniform access to data, software and knowledge. In: Clayton P.D., eds. *Proceedings of the 15th Symposium on Computer Applications in Medical Care*. Washington DC. McGraw-Hill, New York. 1991:496-500.
- [8] Van Mulligen E.M., Timmers T., Van Bommel J.H. A new architecture for integration of heterogeneous software components. *Meth Inform Med* 1993;32:292-301.
- [9] Van Mulligen E.M. An architecture for an integrated medical workstation; its realization and evaluation. 1993, PhD-thesis.
- [10] Timmers T., Van Mulligen E.M., Van den Heuvel F. Integrating clinical databases in a medical workstation using knowledge-based modelling. In: Lun K.C., Degoulet P., Piemme T.E., Rienhoff O., eds. *Proceedings of the Seventh World Congress on Medical Informatics*, Geneva. North-Holland, Amsterdam, 1992:478-82.
- [11] Van den Heuvel F., Timmers T., Van Mulligen E.M., Hess J. Knowledge-based modeling for the classification and follow-up of patients with congenital heart disease. *Proceedings of Computers in Cardiology Conference*. Venice, 1991. IEEE Computer Society Press, Los Alamitos. 1991:737-40.
- [12] Shortliffe E.H.. Computer programs to support medical decisions. *JAMA* 258:61-66.
- [13] Wyatt J., Spiegelhalter D. Evaluating medical decision-aids: what to test, and how. In: Talmon J., Fox J., eds. *System Engineering in medicine*. Springer Verlag, Heidelberg, 1989. pp1-13.
- [14] Van Mulligen E.M., Timmers T., Van Bommel J.H. User Evaluation of an integrated medical workstation for clinical data analysis. (accepted for *Meth Inform Med*).
- [15] Bloem R.M., Zwavelink A., Stijnen Th. Adenocarcinoma of the colon and rectum: a report on 624 cases. *The Netherlands Journal of Surgery*. 1988, 40 (5):121-6.